Statistische und stochastische Grundlagen

Vorlesungsmitschrieb zum Modul an der Universität Stuttgart

Simon König

Inhaltsverzeichnis

1

Kapitel 1: Statistik

1: Themenbereiche

1.1 Desktriptive Statistik

1.2 Explorative Statistik

1.3 Induktive Statistik

2: Grundbegriffe

2.1 Grundbegriffe der Statistik

Statistische Einheit Objekte die erfasst werden und an denen die interessierenden Größen erfasst werden

Grundgesamtheit Menge aller für die Fragestellung relevanten statistischen Einheiten

Teilgesamtheit Teilmenge der Grundgesamtheit

Stichprobe Tatsächlich untersuchte Teilmenge der Grundgesamtheit

Merkmal, Variable Größe von Interesse

Merkmalsausprägung, Wert Konkreter Wert des Merkmals für eine bestimmte statistische Einheit

2.2 Charakterisierung der Merkmale

diskret Merkmale, die nur endlich viele oder abzählbar unendlich viele Ausprägungen annehmen sind diskret.

stetig Merkmale, die Werte aus einem Intervall annehmen können heißen stetig.

quasi-stetig Merkmale, die sich nur diskret messen lassen aber aufgrund einer sehr feinen Abstufung wie stetige Merkamle behandelt werden können.

Die Ausprägungen eines stetigen Merkmals lassen sich immer so zusammenfassen, dass es als diskret angesehen werden kann. Die Ausprägungen heißen dann gruppiert oder klassiert.

2.3 Skalen

Zusätzlich zur Charakterisierung der Merkmale werden diese anhand ihres Skalenniveaus unterschieden.

Nominalskala Wenn die Ausprägungen Namen oder Kategorien sind, die den Einheiten zugeordnet werden heißt das Merkmal *nominalskaliert*. Beispielsweise Geschlecht oder Verwendungszweck.

Ordinalskala Merkmale mit Ausprägungen zwar mit Ordnung, bei denen allerdings ein Abstand der Merkamale nicht interpretier- oder vergleichbar ist heißen *ordinalskaiert*. Ein Beispiel hierfür wären Schulnoten.

Kardinalskala Ein kardinalskaliertes Merkmal wird oft auch metrisch bezeichnet. Hierbei sind die Abstände der Ausprägungen interpretierbar und zusätzlich ist ein sinnvoller Nullpunkt der Skala festgelegt oder bestimmbar.

Auf Basis dieser Skalenmerkmale nennt man Merkmale mit endlich vielen Ausprägungen, die höchstens ordinalskaliert sind *qualitative* oder *kategoriale Merkmale*. Diese geben eine Qualität aber nicht ein Ausmaß wieder.
Geben die Ausprägungen jedoch eine Intensität oder Ausmaß wieder so spricht man von *quantitativen Merkmalen*. Alle Messungen mit Zahlenwerten stellen Ausprägungen quantiativer Merkmale dar. Ein kardinalskaliertes Merkmal ist stets quantitativ.

2.4 Datengewinnung, Datenerhebung

S.18 ff

3: Verteilungen und ihre Darstellungen

3.1 Häufigkeiten

Als *Urliste* bezeichnet man die Menge der Merkmale X der Untersuchungseinheiten $U = \{x_1, \ldots, x_n\}$. Die *auftretenden Ausprägungen* von X sind die Werte $\{a_1, \ldots, x_n\} \subseteq \{x_1, \ldots, x_n\}, k \leq n$. Oftmals treten in einem großen Datensatz der Größe n nicht auch n verschiedene Werte x_i auf.
Damit definieren sich

Definition 3.1: Absolute Häufigkeit

Die absolute Häufigkeit einer auftretenden Ausprägung a in einer Urliste U ist

$$h(a) = |\{x \in U \mid x = a\}|\,.$$

Es gilt immer, dass die Summe aller absoluten Häufigkeiten gleich der Datensatzgröße ist

$$\sum_{i=1}^{n} h(a_i) = |U|.$$

Die absolute Häufigkeitsverteilung ist dargestellt durch die Folge von Werten

$$h_1, \ldots, h_k = h(a_i), \ldots, h(a_k)$$

Definition 3.2: Relative Häufigkeit

Die relative Häufigkeit einer auftretenden Ausprägung a in einer Urliste U ist

$$f(a) = \frac{h(a)}{|U|}.$$

Es gilt ähnlich wie bei der absoluten Häufigkeit für die Summe

$$\sum_{i=1}^{n} f(a_i) = 1.$$

Eine grafische Darstellung einer Häufigkeitsverteilung nennt man ein *Histogramm*. Bei Histogrammen ist auf die Flächentreue zu achten, das bedeutet, dass der Flächeninhalt der aufgetragenen Rechtecke proportional (oder gleich) zu h_j oder f_j ist. So kann das menschliche Auge die Verteilung besser wahrnehmen.
Hat das Histogramm einer Verteilung nur einen deutlich erkennbaren Hochpunkt (Gipfel), heißt sie *unimodal*. Treten mehrere Gipfel auf nennt man die Verteilung *multimodal*. Bei zwei Gipfeln spricht man von einer *bimodalen* Verteilung.

Man nennt eine Verteilung *symmetrisch*, wenn es eine Symmetrieachse gibt, sodass die rechte und linke Hälfte der Verteilung annähernd zueinander spiegelbildlich sind. Eine Verteilung heißt *schief*, wenn sie deutlich unsymmetrisch ist. Sie heißt dann *linkssteil oder rechtsschief*, wenn der überwiegende Anteil von Daten linksseitig konzentriert ist. Dann steigt die Verteilung links deutlich steiler ab als rechts. Entsprechend *rechtssteile oder linksschiefe* Verteilungen.

3.2 Kumulierte Häufigkeiten

Die kumulierten Häufigkeitsverteilungen geben an, wie viele Datenpunkte der Urliste, beziehungsweise welcher Anteil der Daten unterhalb einer Schranke liegen. Um diese Aussage sinnvoll zu beantworten ist zumindest eine Ordinalskala nötig.

Definition 3.3: Absolute kumulierte Häufigkeitsverteilung

Die absolut kumulierte Häufigkeitsverteilung ist die Funktion

$$H(x) = \sum_{i:a_i \leq x} h_i.$$

Definition 3.4: Relative kumulierte Häufigkeitsverteilung

Die relative kumulierte Häufigkeitsverteilung oder auch empirische Verteilungsfunktion ist

$$F(x) = \sum_{i:a_i \leq x} f_i.$$

Die kumulierten Häufigkeitsverteilungen sind monoton wachsende, Treppenfunktionen, die an den Sprungstellen rechtsseitig stetig sind.

3.3 Gruppierung

Sind alle auftretenden Ausprägungen Elemente eines Interfalls $[a, b]$, lässt sich dieses in gleich große Klassen der Größe d unterteilen.
Eine Klassifizierung ist allgemein

$$[a, c_1), \ldots, [c_i, c_{i+1}), [c_{i+1}, c_{i+2}), \ldots \quad \forall i : c_{i+1} - c_i = d$$

Klassifizierte Daten sind i.A. einfacher zu interpretieren als große Mengen von Daten, die sich nur wenig voneinander unterscheiden.
Der maximale Fehler bei der Klassifizierung ist die halbe Klassengröße.

3.4 Lagemaße

Lagemaße helfen beim Vergleich verschiedener Eigenschaften, bzw dem Vergleich verschiedener statistischer Einheiten mit einer gemeinsamen Eigenschaft.

Definition 3.5: Lagemaß

Ein *Lagemaß* ist eine Abbildung $L : \mathbb{R}^n \to \mathbb{R}$ mit der Eigenschaft

$$L(x_1 + a, \ldots, x_n + a) = L(x_1, \ldots, x_n) + a \quad \forall a, x_i \in \mathbb{R} \ (1 \leq i \leq n)$$

Ein Lagemaß beschreibt das Zentrum einer Verteilung.

Beispiele für Lagemaße sind

3.4.1 Arithmetisches Mittel

Das arithmetische Mittel ist nur für quantitative Merkmale sinnvoll. Es berechnet sich durch

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (a_i \cdot f_i)$$

aus Rohdaten beziehungsweise aus den Häufigkeitsdaten.
Mit dem arithmetischen Mittel gilt die sogenannte *Schwerpunkteigenschaft*

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

Wie aus der Formel erkennbar, ist das arithmetische Mittel extrem empfindlich gegen Ausreißer. Dafür wurden die folgenden Mittel eingeführt.

Das getrimmte Mittel Um Ausreißer weniger stark ins Gewicht fallen zu lassen wird der Datensatz absichtlich verkleinert. Beim getrimmten Mittel aus einer sortiert vorliegenden Liste von Daten werden zum Beispiel die oberen und unteren 5% der Daten abgeschnitten, damit fallen auch eventuelle Ausreißer raus. Die Datensatzgröße bleibt jedoch nicht erhalten.

Das winsorisierte Mittel Ähnlich wie beim getrimmten Mittel wird der Datensatz beim winsorisierten Mittel von oben und unten herein bearbeitet. Anstatt Daten zu löschen werden beispielsweise die oberen 5% durch den nächstkleineren Wert ersetzt. Hierbei bleibt also die Datensatzgröße gleich.

3.4.2 Median

Der Median stellt ein robusteres Lagemaß als das arithmetische Mittel dar, er ist resistenter gegen Ausreißer im Datensatz. Für $x_1 \leq x_2 \leq \ldots \leq x_n$, also einen sortiert vorliegenden Datensatz ist der Median

$$x_{\text{med}} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ gerade} \end{cases}$$

Benötigt eine Zahlenordnung, also eine Ordinalskala.
Mindestens 50% der Daten sind kleiner oder gleich x_{med}, genauso sind mindestens 50% der Daten größer oder gleich dem Median x_{med}.

3.4.3 Modus

Der Modus ist die Ausprägung größter Häufigkeit $x_{\text{mod}} = a_i$ mit $h(a_i) = \max\{h(a) \mid a \in A\}$ wobei A die Menge aller vorkommenden Ausprägungen der Urliste ist. Der Modus ist dann eindeutig, wenn die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.
Der Modus empfiehlt sich schon für nominalskalierte Daten.

3.4.4 Geometrisches Mittel

Für eine Urliste $U = \{u_1, \dots, u_n\}$ ist das geometrische Mittel definiert als

$$x_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} u_i}.$$

Das geometrische Mittel wird z.B. bei der Berechnung des effektiven Jahreszinses verwendet, es stellt jedoch kein Lagemaß im engeren Sinne dar.

3.4.5 Harmonisches Mittel

Für eine Urliste $U = \{u_1, \dots, u_n\}$ ist das harmonische Mittel definiert als

$$x_{\text{harm}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}.$$

Genauso wie das geometrische Mittel zählt das harmonische nicht zu den Lagemaßen im engeren Sinne.

3.5 Lageregeln

Für symmetrische Verteilungen gilt $\overline{x} \approx x_{\text{med}} \approx x_{\text{mod}}$.
Für linkssteile Verteilungen $\overline{x} > x_{\text{med}} > x_{\text{mod}}$.
Und ebenso für rechtssteile Verteilungen $\overline{x} < x_{\text{med}} < x_{\text{mod}}$.

Literaturverzeichnis

[1] L. Fahrmeir, *Statistik: Der Weg zur Datenanalyse*, Springer-Lehrbuch : SpringerLink : Bücher, Springer Spektrum, Berlin, Heidelberg, 8. aufl. 2016 ed., 2016.

[2] N. Henze, *Stochastik für Einsteiger: Eine Einführung in die faszinierende Welt des Zufalls*, SpringerLink : Bücher, Springer Spektrum, Wiesbaden, 10., überarb. aufl. 2013 ed., 2013.

[3] U. Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg Studium, Aufbaukurs Mathematik : SpringerLink : Bücher, Vieweg+Teubner Verlag, Wiesbaden, 8., erweiterte auflage ed., 2005.